The University Of Sheffield.

BMS 353
Bioinformatics for Biomedical Science

Module coordinator: Dr Marta Milo
Research Software Instructor: Dr Mike Croucher

The University Of Sheffield.

Today's Outline

Part A :Presentation of the module

Break – question answering

Part B :Introduction

The University Of Sheffield.

Part A
Presentation of the module

The University Of Sheffield.

What is all about?

This module aims to provide an range of the fundamental concepts and technologies underlying computational biology and bioinformatics.

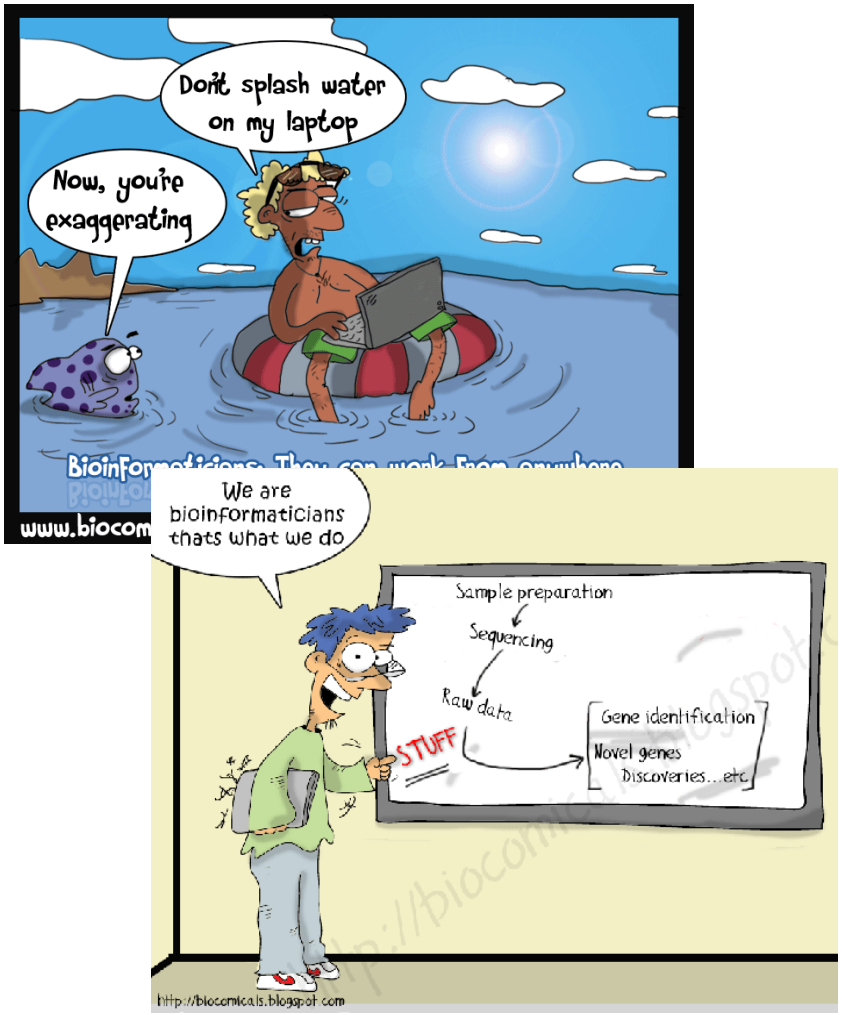
Computational Biology *is the development and application of data driven mathematical modeling and computational simulation techniques to study of biological, behavioral, and social systems*

Bioinformatics *is an interdisciplinary field of science that develops methods and software tools for understanding biological data Bioinformatics combines computer science, statistics, mathematics, and engineering to analyse and interpret biological data.*

adapted from wikipedia

The University Of Sheffield.

What is a Bioinformatician?



What I do in my research:

- Next generation Sequencing data analysis
- Noise deconvolution
- Modelling uncertainty
- Integration of data
- Modelling observed data for predictions

What am I going to teach you?

Some of that STUFF

The University Of Sheffield.

What are the learning outcomes of this module?

This module aims to:

1. provide an understanding of the *fundamental concepts* and technologies underlying computational biology and bioinformatics

2. it is aimed at biology students with basic knowledge of *mathematical concepts*, to equip them with methods of Bioinformatics and Computational biology for the analysis of biological data

3. provide skills in experimental design and data pipeline generation

4. it will use a *multidisciplinary* approach integrated with programming tools and statistical concepts underpinning advanced data analysis and methods that are suitable *for high-throughput data analysis*

5. provide new transferable skills

The University Of Sheffield.

How will you be learning?

- Lectures on theoretical concepts
- Online resources from open source software
- Writing simple scripts for data analysis during practical classes
- Small research project on real data
- Group discussion and forum through the module website
- Banging your head on the computer ..
- Giving yourself time to adapt to this new way of thinking…

The University Of Sheffield.

What will you gain from BMS353?

- Training in data analysis and basic programming skills with the aims of a) awareness of the effects of experimental design in the subsequent data analysis
- A good understanding of technologies and methods for Bioinformatics and use of workflow and pipelines for data analysis
- New qualifications that will increase your employability
- Deeper insight into the principles of conducting a research data analysis project
- A new se of transferable skills, like programming and awareness of cloud computing and data sharing
- Learning a new terminology and new interdisciplinary skills

The University Of Sheffield.

Module Outline

The teaching consists of two hours of lectures and two of lab classes each week. The lectures are on Thursdays, the labs on Fridays. The teaching schedule and venue for each week are given below:

Lectures:

Wk 7 on Thur 12Nov	3.00-5.00pm	in **HI-G25**	*Introduction of the course and tools*
Wk 8 on Thur 19Nov	3.00-5.00pm	in **HI-G25**	*Concepts of statistics and their implementation for data analysis.*
Wk 9 on Thur 26Nov	3.00-5.00pm	in **HI-G25**	*Bioinformatics for high throughput data.*
Wk 10 on Thur 03Dec	3.00-5.00pm	in **HI-G25**	*Bioconductor and limma*
Wk 11 on Thur 10Dec	3.00-5.00pm	in **HI-G25**	*Functional Annotation & pathway analysis.*
Wk 12 on Thur 17Dec	3.00-5.00pm	in **HI-G25**	*Project Allocation and Guest Lecture*

The University Of Sheffield.

Module Outline (cont.)

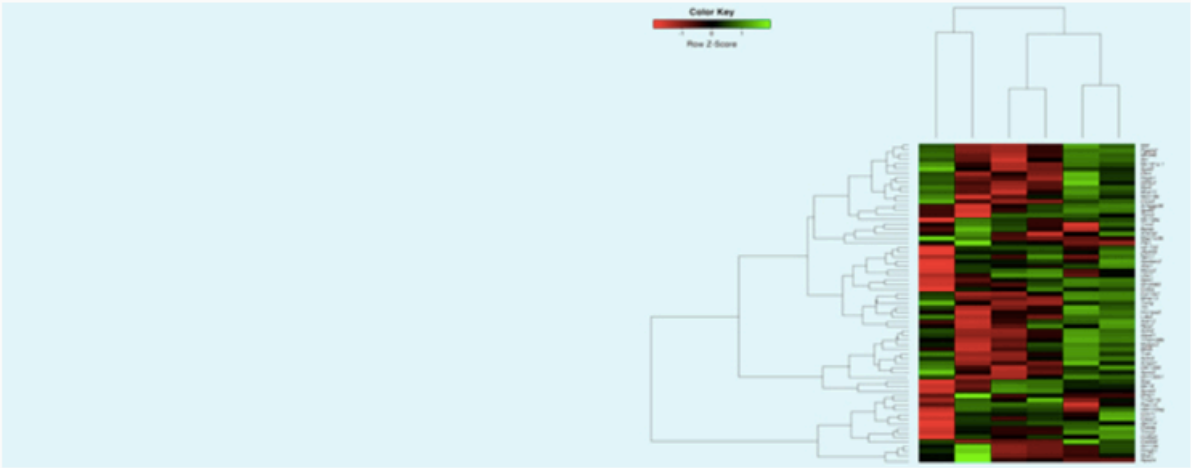
Practical classes:

Wk 7 on Fri 13Nov	3.00-5.00pm	in **IC 3.02**	*Basics of data manipulation using scripts*
Wk 8 on Fri19Nov	3.00-5.00pm	in **IC 3.02**	*Basics of data analysis using scripts*.
Wk 9 on Fri 27Nov	3.00-5.00pm	in **IC 3.02**	*Bioinformatics analysis for high throughput data*
Wk 10 on Fri 04Dec	3.00-5.00pm	in **IC 3.02**	*Analysis of high throughput data with Bioconductor and limma*
Wk 11 on Fri 11Dec	3.00-5.00pm	in **IC 3.02**	*Implementation of Functional and Pathways Analysis*
Wk 12 on Fri 18Dec	3.00-5.00pm	in **IC 3.02**	*Experimental design and projects pipelines. Wrap up.*

The University Of Sheffield.

BMS353 web site

BMS353 Bioinformatics for Biomedical Scientists



Deadlines Exam Feedback About Overview

Overview

Overview

BMS353 - Bioinformatics for Biomedical Science Module Co-ordinator: Dr M Milo Research Software Instructor: Dr Mike Croucher Semester 1B – Level 3 Biomedical Science 2015/2016

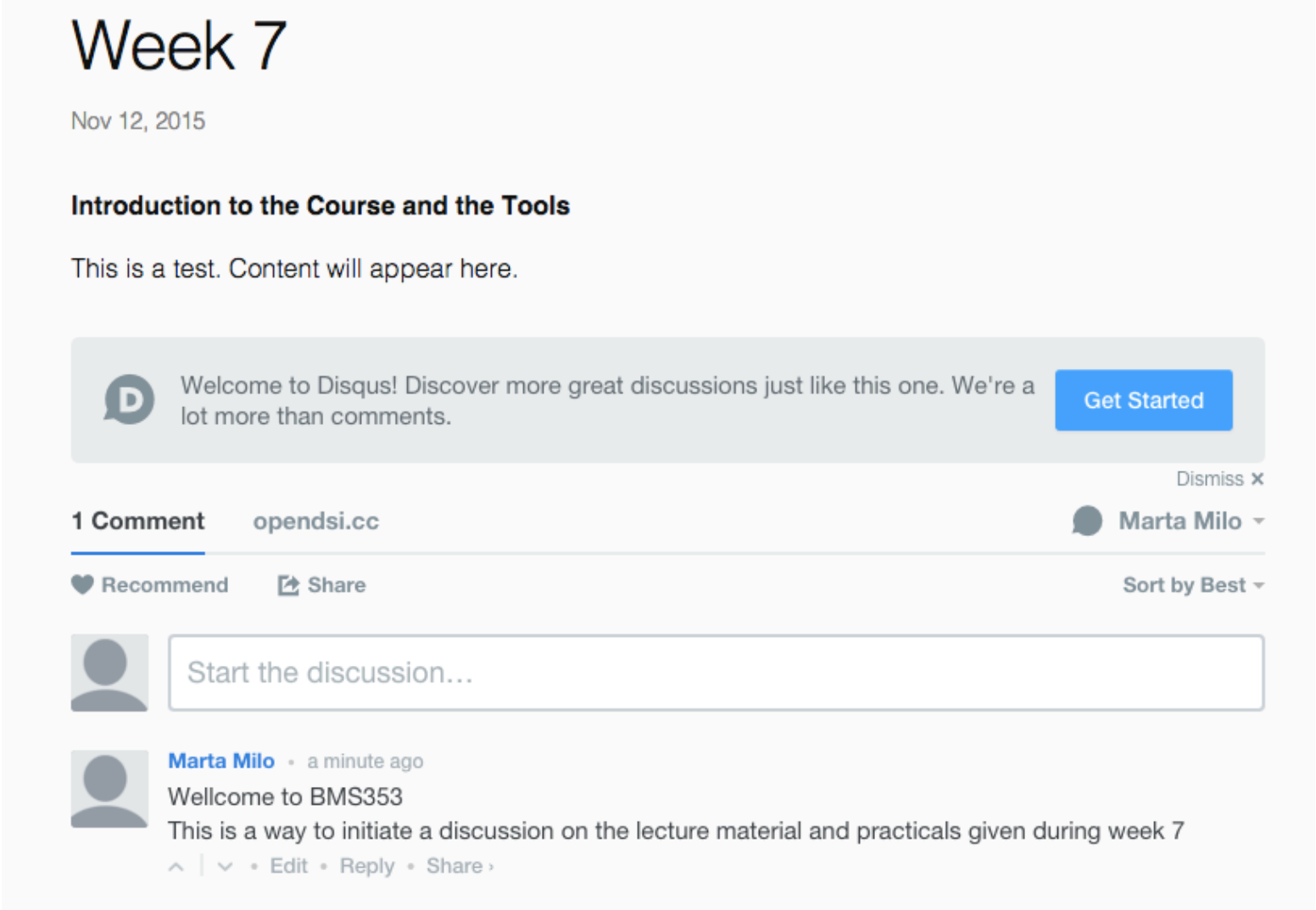
Module Overview

This module aims to provide an understanding of the fundamental concepts and technologies

The University Of Sheffield.

BMS353 discussion forum

Discussion and feedback will happen via this channel and shared



The University Of Sheffield.

The tools we will use

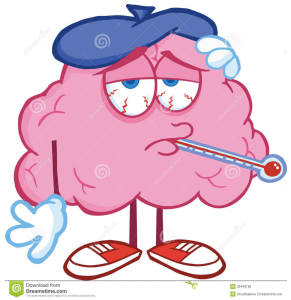
Jupyter notebook (Originally Ipython notebook)

Combines computer programs (code), text, data, results into one interactive document

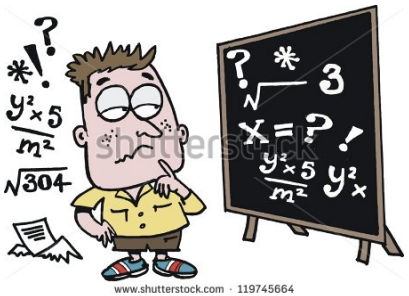
SageMathCloud™

We will use a cloud computing environment called SageMathCloud

A popular programming language in areas such as bioinformatics, statistics and data analysis.



We will use our brain to create knew knowledge



Some mathematical concepts

The University Of Sheffield.

BMS353 assessment

The exam for this module will be split in two parts:

Part A – A Multiple Choice Question test for the duration of 1hr and 30 minutes, that will count 30% of the final grade

Part B – A notebook with the implementation of allocated projects that will count for 70% of the final grade.

The project will be a collection of all the tools experienced in the practical labs implemented on a set of real data. It will be developed in groups of three students, but notebook will have to be handed individually.

The lab practical notebooks handed in every week during the module will constitute formative feedback that can be used for the final project.

The University Of Sheffield.

MCQ assessment

Each question will have 4 possible responses A, B, C or D. **ONLY ONE RESPONSE IS CORRECT IN EACH CASE.** Each question is worth one mark, correct answer will count as 1, an incorrect answer will count as -0.5. **Not answered questions will count as 0**.

1. What is the main subject of BMS353:
A. Phycology
B. Statistics
C. Computational biology
D. Computer Science

2. What level students BMS353 is aimed at:
A. Level 3 - BMS
B. Post graduate
C. Master students
D. Computer Science students

3. There will be no mathematics in BMS353:
A. TRUE
B. FALSE
C. TRUE only in odd days
D. TRUE only in even days

Student 1: 1C, 2B, 3A mark=0

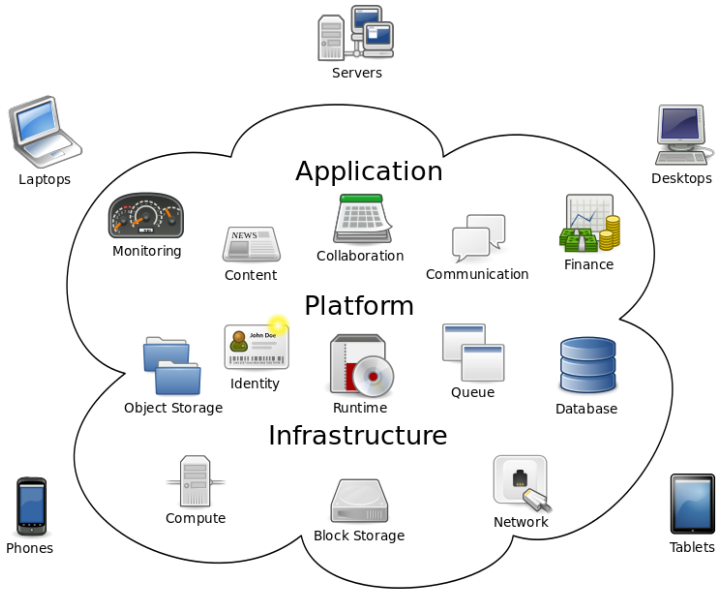
Student 2: 1C, 2-, 3- mark=1

The University Of Sheffield.

Part B
Introduction

Cloud computing

Cloud computing, or simply "*the cloud*", also known as on-demand computing is a model for enabling on-demand access to a shared pool of configurable resources



Cloud Computing
Sam Johnston – from Wikipedia

The cloud metaphor: the network elements representing the services are invisible to the user, like obscured by a cloud

Advantages

- Cost efficient
- Large space storage
- Backup and recovery
- Easy access
- Quick to gain functionality
- Incentives collaboration and data sharing

Disadvantages

- Technical issues
- Security in the cloud
- Prone to attack

The University Of Sheffield.

Cloud computing: an example

A very effective use of the cloud resources and its commercial exploitation is given by Amazon

They used cloud computing to create the concept of **Elastic Computing (EC2).**
It is a key part of the Amazon Web Services (AWS), which is composed of scalable elastic compute unit (ECU) that were introduced as an abstraction of computer resources.

A user can create, launch, and terminate server usage as needed. It is based on a "paying by the hour for active servers" this is why it is called "elastic".

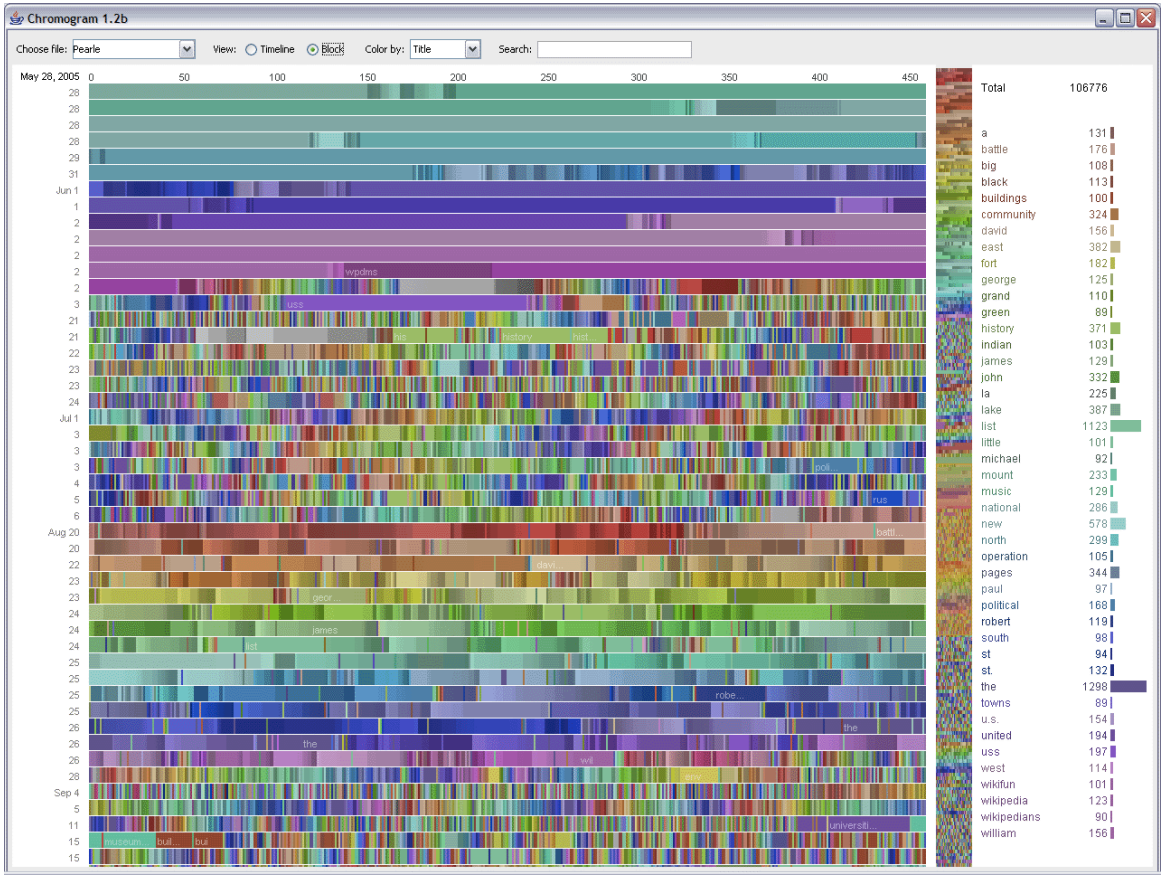
Its global feature allows users to control over the geographical location of instances (server usage), optimising latency and redundancy.

First to allow company to rent scalable computing resources
Their retail ecommerce site is entirely base on cloud computing

The University Of Sheffield.

Big Data and Data Sharing

Big data is a very generic term to indicate datasets that are so large or complex that traditional data processing applications are inadequate for mining it.



Visualization of daily Wikipedia edits created by IBM. At multiple terabytes in size, the text and images of Wikipedia are an example of big data.

High volume
High velocity
High variety
Highly variable
High variation in quality
High complexity

The University Of Sheffield.

Big Data and Data Sharing (cont.)

There are many challenges when dealing with big data, some of them are:

- Data analysis
- Data curation
- Searching engines
- Data sharing
- Data storage and transfer
- Data visualization
- Information privacy

However, big data has a very *predictive power* and its accuracy may lead to more confident decision making.

In biology:

With the advent of high-throughput genomics, life scientists are starting to grapple with massive datasets, encountering big data challenges

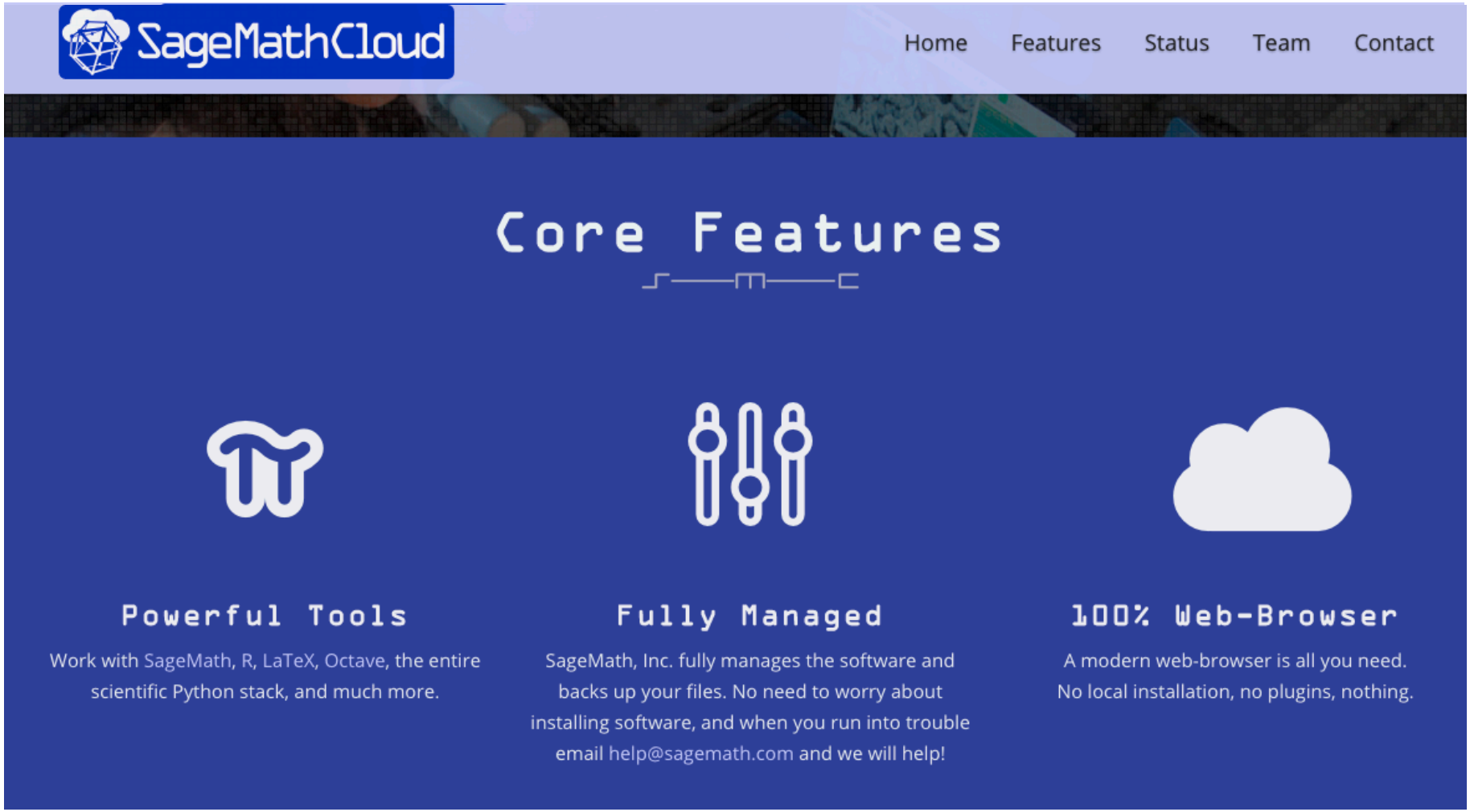
Technology Feature, Nature 2013

Analysing the large amount of genomic data with local infrastructure is impossible.
The data is then moved to the cloud for analysis and storage.
Data sharing is becoming crucial for biological data.

The University Of Sheffield.

SageMathCloud

www.sagemath.com



The University Of Sheffield.

Jupyter Notebooks on SageMathCloud

We will use Jupyter Notebooks and their kernels on SageMathCloud for all our practical classes .

A Jupyter Notebooks **kernel** is a "*computational engine*" that executes the code written in the Notebook document.

In this module (BMS353) we will use R kernels to implement our data analysis in the notebooks.

There will be allocated folder and storage space to our *project:* BMS353
You will access your assignments and data using SageMathCloud with a web browser.

Everything will be stored in SageMathCloud *folder* allocated to you.
The cloud will backup and secure our work, as well as giving us computational time for the data anlysis

All the lab practicals and the final project will be marked and assessed from notebooks saved in the SageMathCloud *folders.*

The University Of Sheffield.

Basic programming terminology

Programming language = is a language formally designed to communicate instruction to a machine, i.e. a computer, to control behavior or to express a mathematical construct in numerical form (make operations, more or less complex)

Algorithm = it is a procedure or formula for solving a problem

Kernels = computational engine that is activated by a specific language (i.e. R, Python, C etc.)

Scripts = a list of instructions that represent the command needed to represent a task. It has a logical structure and a defined structure for data input

Implementation = the process of putting into effect the list of instructions that are specified in the script. This is done by using numerical values as input. The implementation process will produce a a final set of values.

Debug = Process for identifying and removing error from scripts

Object = virtual container of values stored in the working space. It is used to implement the instructions and to store values during the implementation and as final set.

Programming Function = it is a procedure or a routine that encapsulate a "task". Many instructions are combined in one "word" (the name of the programming function) which will implement that "task" on a set of specified input.

Read and Write = The process of uploading data into the work space and to download data from the working space into a local or remote archive (folder)

The University Of Sheffield.

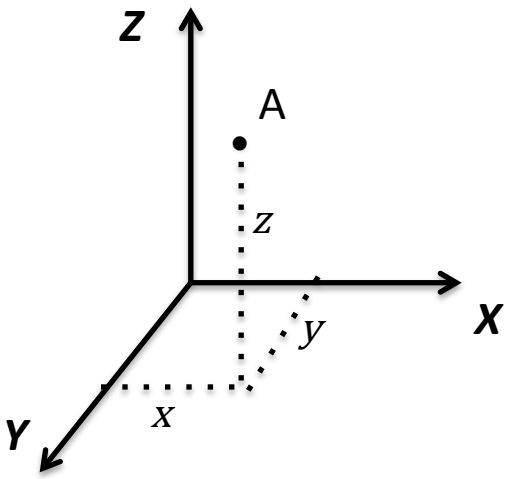
Basic mathematics

Basic mathematics notation

Single values and vectors

$$x, y \in \Re$$

x and y are values from the real numbers



$$\overline{A} \equiv (x, y, z)$$

In general

$$\overline{x} \equiv (x_1, x_2, \ldots, x_N)$$

$$x_i \in \Re$$

$$i = 1, \ldots, N$$

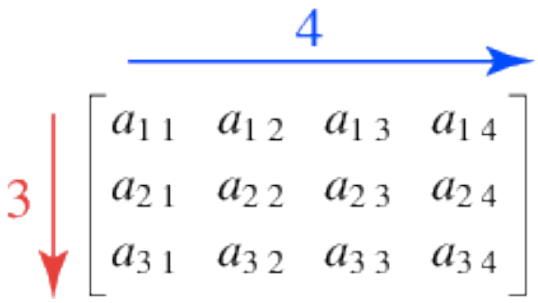
The values x_i are called variables since they can assume a range a fixed values
The parameter are fixed values that we indicate in mathematical notation with Greek letters

$$\alpha, \beta, \mu, \sigma, \lambda \ldots\ldots$$

Basic mathematics notation (cont.)

Matrices are tables of values or letters that are organised in rows and columns. In common use they only have two dimensions, in more advanced use they can have three.

Vectors are special cases of matrices, they have a number of *N* columns and only one row



$$A = [3x4]$$

3×4 matrix

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Operation with Matrix

Sum and Difference same dimensions

Multiplications number of column of the first matrix need to be the same as number of raw of the second matrix. Multiplication is done so that:

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} = 1 \cdot 4 + 2 \cdot 5 + 3 \cdot 6 = 22$$

$$\overbrace{1 \times 3} \quad \overbrace{3 \times 1} \quad = \quad \overbrace{1 \times 1}$$

The University Of Sheffield.

Basic mathematics notation (cont.)

A way of writing a notation for large sums or multiplication is to use the Greek symbols of

$\sum$ For summing

$\prod$ For multiplying

For summing *N* values we will use the following notation: $\sum_{i=1}^{N} x_i / \sigma$

For multiplying *N* values we will use the following notation: $\prod_{i=1}^{N} x_i / \sigma$

Basic mathematics notation (cont.)

A function is a relation from a set of input to a set of possible outputs, where each input is related to exactly one output.

$$f(x) = x/2$$

output

Input (variable)

$$f(x) = x^4 + 4$$

When the input is one we say a *one-dimension function*

$$f(x,y) = x^2 + y^2$$

When the input is more that one variable we say a *multi-dimension function.* With two variable we say a *bi-dimensional function*

$$f(x,y/\alpha) = \frac{x^2 + y^2}{\alpha}$$

We can also have function *conditional* to a parameter. In this case we call them *conditional functions*

Where α has value from a set of even number between 0 and 10

The University Of Sheffield.

Summary

- What is BMS353 about and what you expect to learn and gain after taking BMS353
- How to gain information about the module and where to find links to additional reading material, lectures content and practical classes (Web site)
- How to interact for discussion and problem-solving
- How you will get assessed
- Tools we will be using in BMS353
- Cloud computing and Big Data
- Jupyter Notebooks and SageMathCloud
- Basic programming terminology
- Refreshed some basic mathematical notions and notations.